# Curation of the
# End-of-Term Web Archive

Kathleen Murray – Lauren Ko – Mark Phillips

IS&T Archiving Conference – May 2011 – Salt Lake City

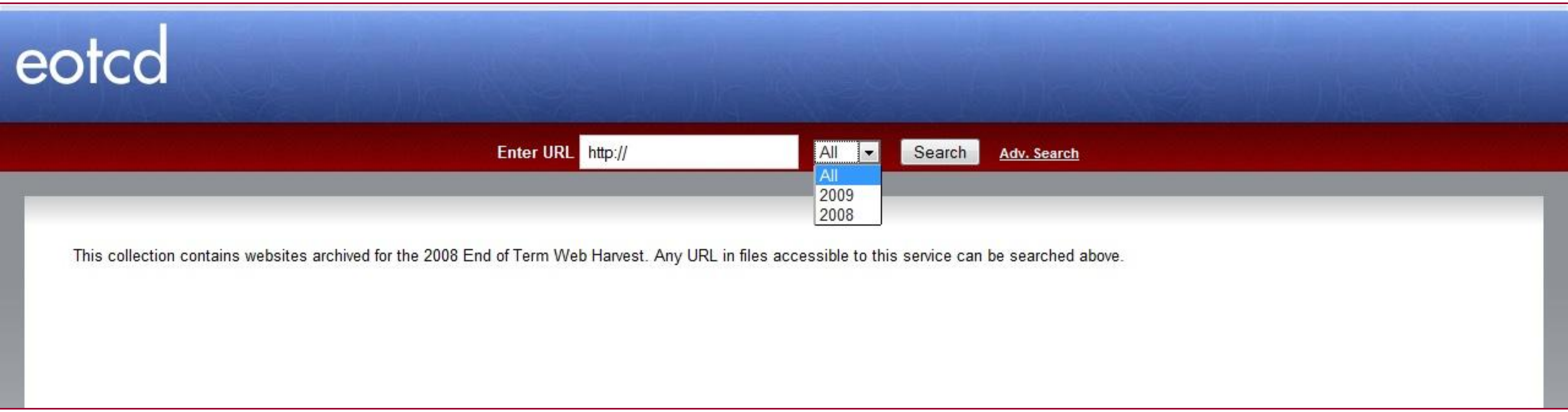
UNT Libraries THE POWER OF IDEAS STARTS HERE IMLS Award LG-06-09-0174-09

# Background: EOT Web Archive

- Who
  - Library of Congress, the GPO, the Internet Archive (IA), the University of North Texas (UNT) Libraries, and the California Digital Library (CDL)
- What
  - Entirety of the federal government's public Web presence
- When
  - Before & after the 2009 change in administrations
- How
  - Nomination Tool: Websites
  - Website Harvests: IA, UNT, & CDL
  - Harvest Consolidation: Library of Congress

# Background: Web Archive Organization

- WARC files (ISO 28500)
  - Specifies formats needed for storage, management, and exchange of data objects (or resources)
  - Applications required to discover and render resources

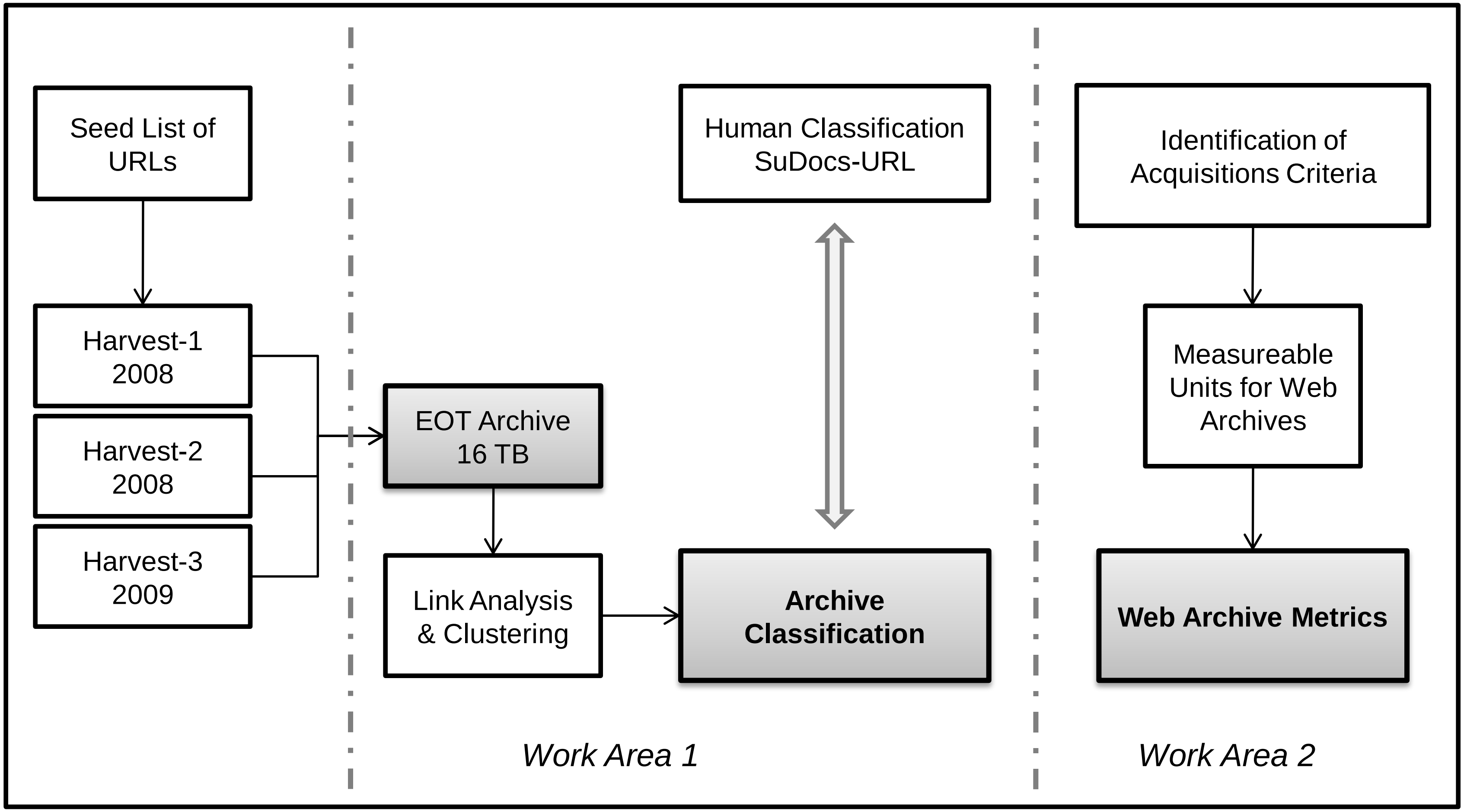


This collection contains websites archived for the 2008 End of Term Web Harvest. Any URL in files accessible to this service can be searched above.

# Background: Problem Statements

- Selection of Materials
  - Foreknowledge of a resource’s URL often required
  - The absence of descriptive metadata or classification schemes thwarts discovery & access
- Metrics
  - Acquisition & retention decisions require standard metrics which are not available

# Background: Work Areas



# CLASSIFICATION

# Classification: Challenges

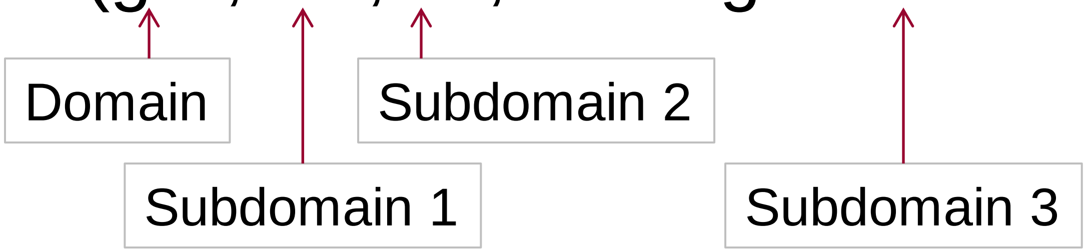
| Largest Domains | # URLs | # Unique Subdomains |
|---|---|---|
| gov | 137,847,822 | 14,339 |
| com | 7,809,711 | 57,873 |
| org | 5,108,645 | 29,798 |
| mil | 3,555,425 | 1,677 |
| edu | 3,552,509 | 13,856 |

Reduced Unique Subdomains to 16,016

# Classification: Managing the Size

SURTS: Reordering URLs by domain structure

Example URL:
http://marriagecalculator.acf.hhs.gov/marriage/
SURT:
http://(gov,hhs,acf,marriagecalculator,)



Domain

Subdomain 1

Subdomain 2

Subdomain 3

Unique Subdomains 1st Level = 1,647
After validation = 1,151 Subdomains

# Human Classification

- SuDocs Classification System
- 10 SMEs classified 1,151 URLs (230/SME)
  - 70% agreement (*n* = 808); 30% disagreement (*n* = 343)
  - Unable to classify: 18 - in scope; 36 - out of scope
- 3 arbitrators classified 343 URLs
  - Assigned SuDocs authors to 286 URLs
  - Unable to classify: 42 - in scope; 15 - out of scope
- Final result:
  - Assigned SuDocs authors to 1,040 subdomains
  - 1,111 authors (1,040 + 71 multiply authored sites)

# Link Analysis: Web Graph

- 1,151 subdomains
  - Multiple URLs per subdomain
  - Example: Library of Congress (LOC) - 44 URLs
    - SURTs format:
      - □ http://(gov,loc,)
      - □ http://(gov,loc,catalog,)
      - □ http://(gov,loc,webarchive,)
- Link extraction: 62,452 links inter-relating HTML files
  - Includes outlinks and inlinks for each URL
- Each pair of linked subdomains assigned a weight
  - Reflecting the number of actual links between the URLs in each source/target subdomain pair

# Cluster Analysis: Clustering Methods

- LinLog Clustering
- Agglomerative Hierarchical Clustering
- Normalized Google Distance (NGD)
- Strongest Outlinks and Majority Inlinks
- Web Communities

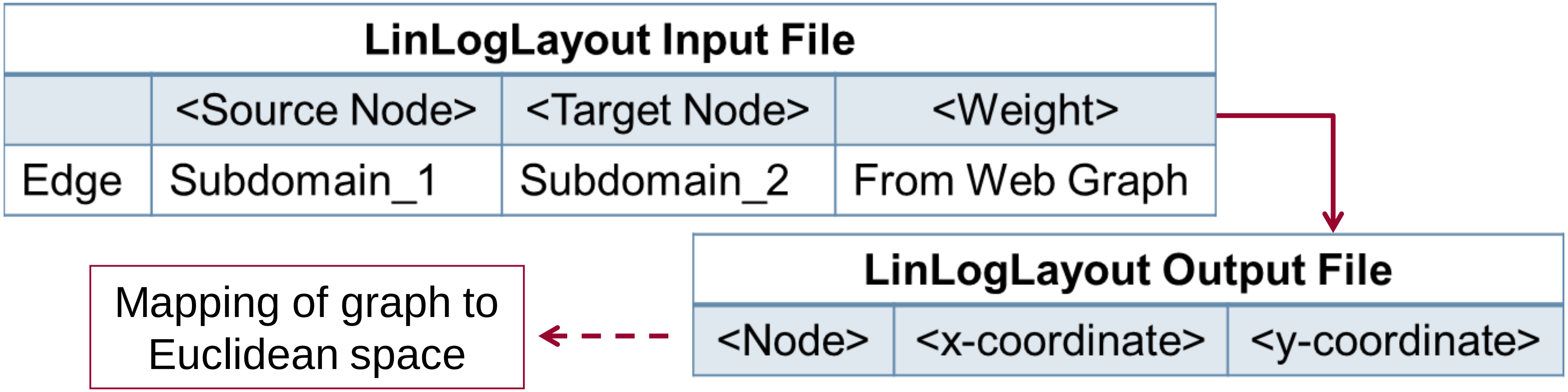
NOTE: Clusters on project wiki: http://research.library.unt.edu/eotcd/wiki/Clusters

# Cluster Analysis: LinLog Clusters

| | Source Node | Target Node | Outlinks | Inlinks |
|---|---|---|---|---|
| **Edge** | Subdomain_1 | Subdomain_2 | # Subdomain_1 | # Subdomain_2 |
| **Edge** | Subdomain_2 | Subdomain_1 | # Subdomain_2 | # Subdomain_1 |

- Two sets of clusters generated
  - 18 node set: Weights on edges = actual number of link occurrences between source & target nodes
  - 20 node set: Weights on edges = ratio of outlinks from a source to a target over all outlinks from that source
- Evaluation
  - Some clusters are larger than expected
  - Ideally a larger number of smaller clusters would result

# Cluster Analysis: Agglomerative Hierarchical Clustering



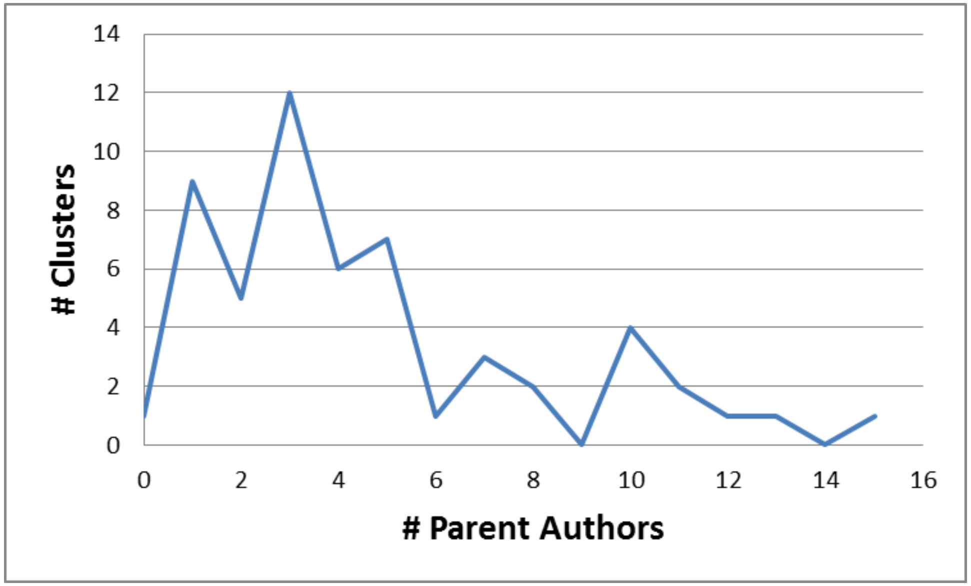
| LinLogLayout Input File | | | |
|---|---|---|---|
| | <Source Node> | <Target Node> | <Weight> |
| Edge | Subdomain_1 | Subdomain_2 | From Web Graph |

| LinLogLayout Output File | | |
|---|---|---|
| <Node> | <x-coordinate> | <y-coordinate> |

Mapping of graph to Euclidean space

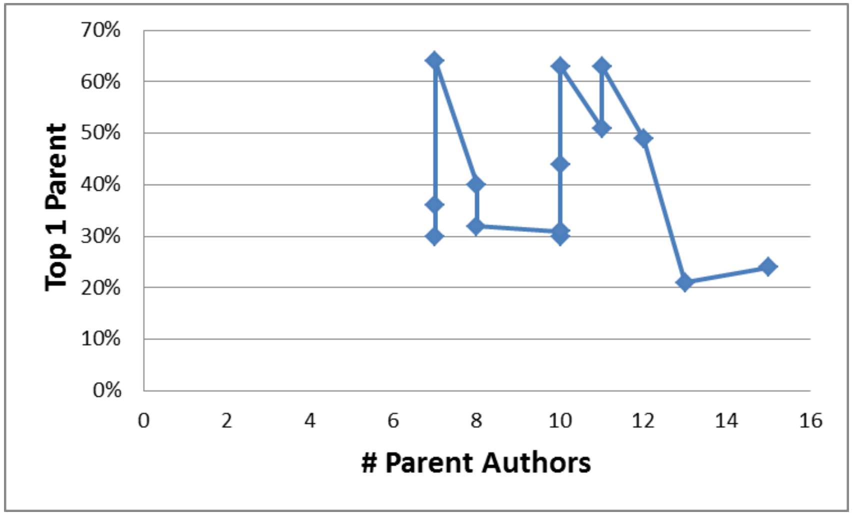
- Two sets of clusters created with groupings set at 55 and 75
- Most successful clustering effort to date; classified both sets using the results of human classification
- Evaluation: Clustering in geometric space is problematic when Web graph is highly linked and its density is highly variable throughout
  - EOT Archive reflects the variances in government agency authors
    - Size; number & size of sub-agencies; amount published

# Findings: Clusters & Parents



- 50% of clusters: ≤ 3 parents
- 75% of clusters: ≤ 6 parents
- 25% of clusters: 7-15 parents

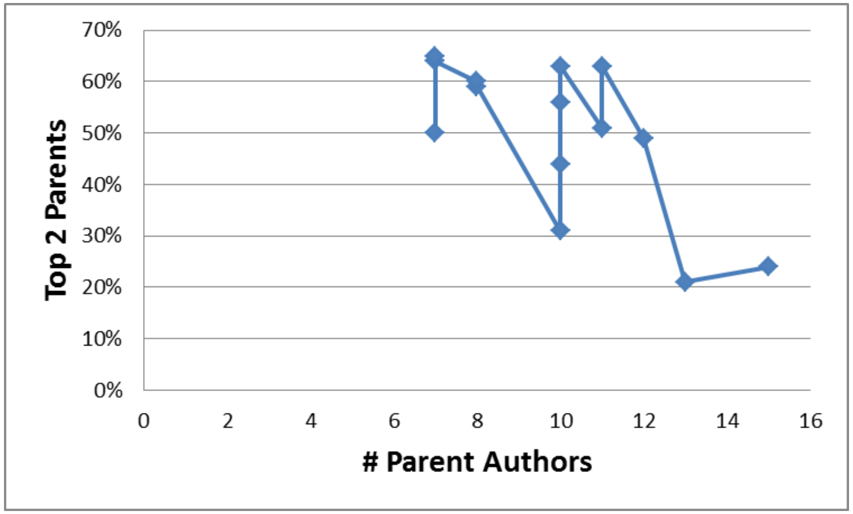
# Findings: Heterogeneity of Parent Authors



14 Clusters: Most heterogeneous

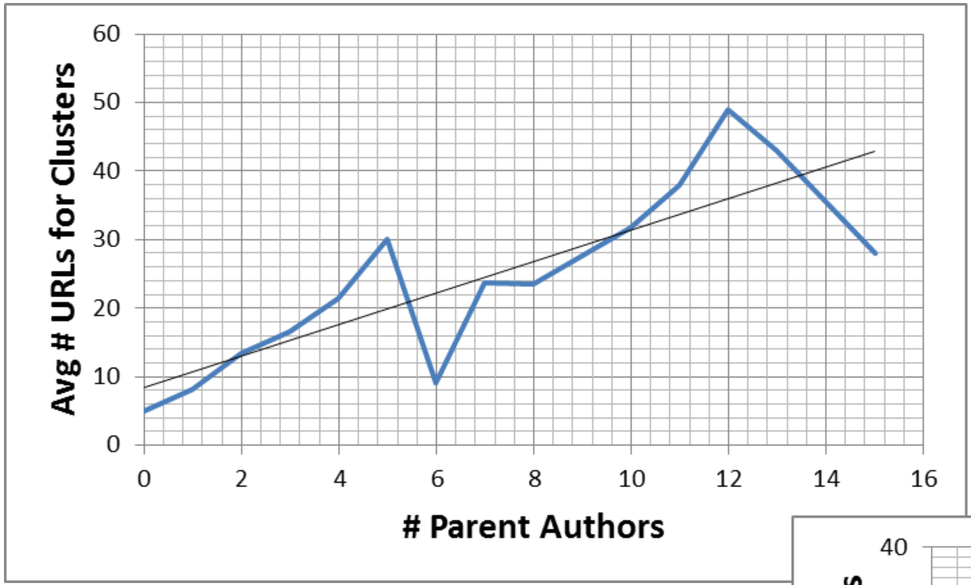
← Five: 1 author 50% or more

Eight: 2 authors 50% or more ↓



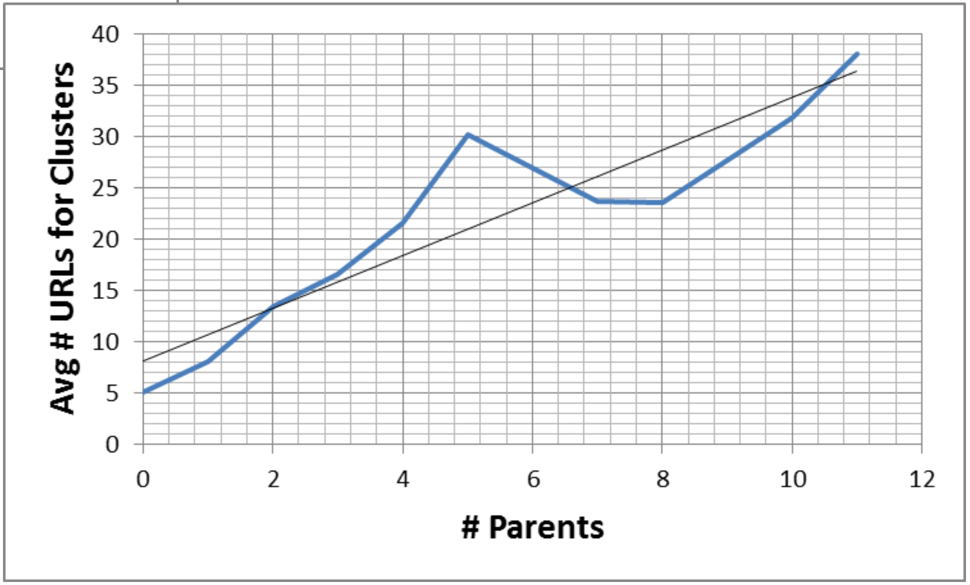
Cluster analysis suggests topical groupings across agency authors

# Findings: Cluster Size & Number of Parents

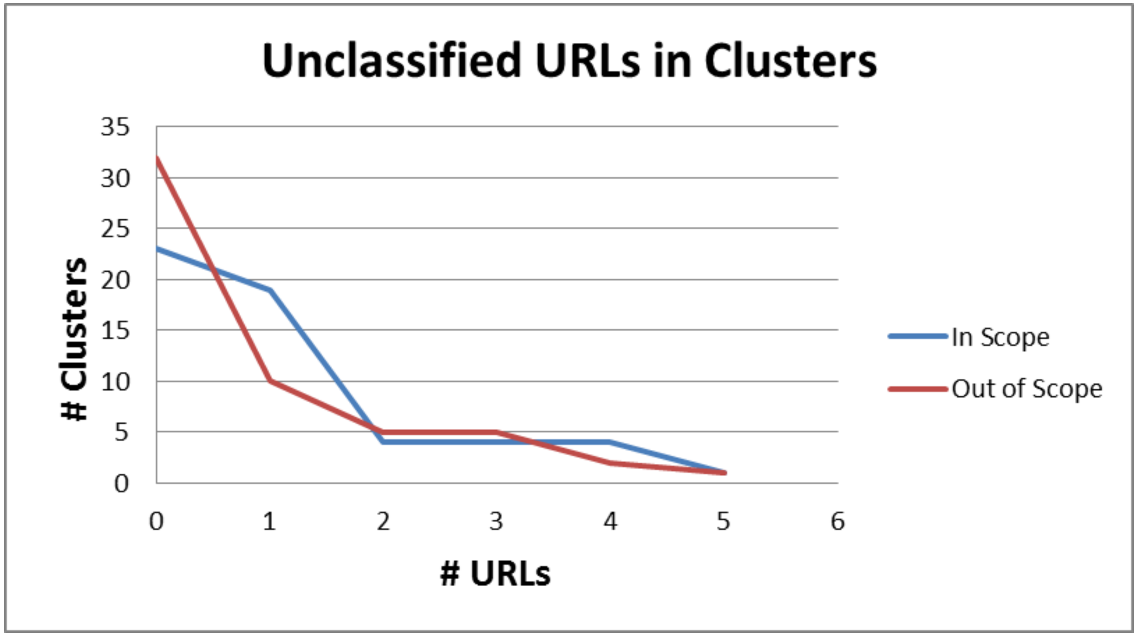


- # Parents = 9 or 14
    - No clusters
- # Parents = 6, 12, 13, & 15
    - 1 cluster each

Suggests that smaller sized clusters might relate to a limited number of SuDoc parent agencies



# Findings: Unclassified URLs



Cluster 4:
Out of Scope

1. dc.gov
2. dcappeals.gov
3. dccourts.gov
4. dcsc.gov
5. washingtondc.gov

Cluster analysis suggests content that falls outside the current classification scheme

# Conclusions

- Involving SMEs in classifying a reasonable sample of a domain-specific Web archive might enable their expertise to be leveraged to:
  - Improve cluster analysis
  - Increase the relevance of search results
- Cluster analysis suggests topical groupings across agency authors
  - Often with 1-2 dominant agency authors
  - Implication for search results:
    - Suggest possible related sites of interest in support of cross-agency subject-related content

# METRICS

# Metrics: Methods

- Focus group discussion with project’s SMEs
  - Identify criteria used for acquisition of materials from Web archives
- Survey of FDLP Libraries
  - Purpose: Assess libraries’ interests and capabilities in accessing v. acquiring content from Web archives
  - Participants: 414 libraries in the Federal Depository Library Program
- Review of current statistics and measurement

# Metrics: Focus Group Findings

- More libraries interested in networked access to an archive v. purchasing and hosting locally
- Current metrics for networked electronic resources are best informants for Web archive content
  - Critical importance of standards compliant usage data
- Authorities - Standards
  - ARL; ACRL; NCES/IPEDS
  - COUNTER: Codes of Practice
    - □ Counting Online Usage of Networked Electronic Resources
  - SUSHI: ANSI/NISO Z39.93-2007
    - □ Standardized Usage Harvesting Initiative

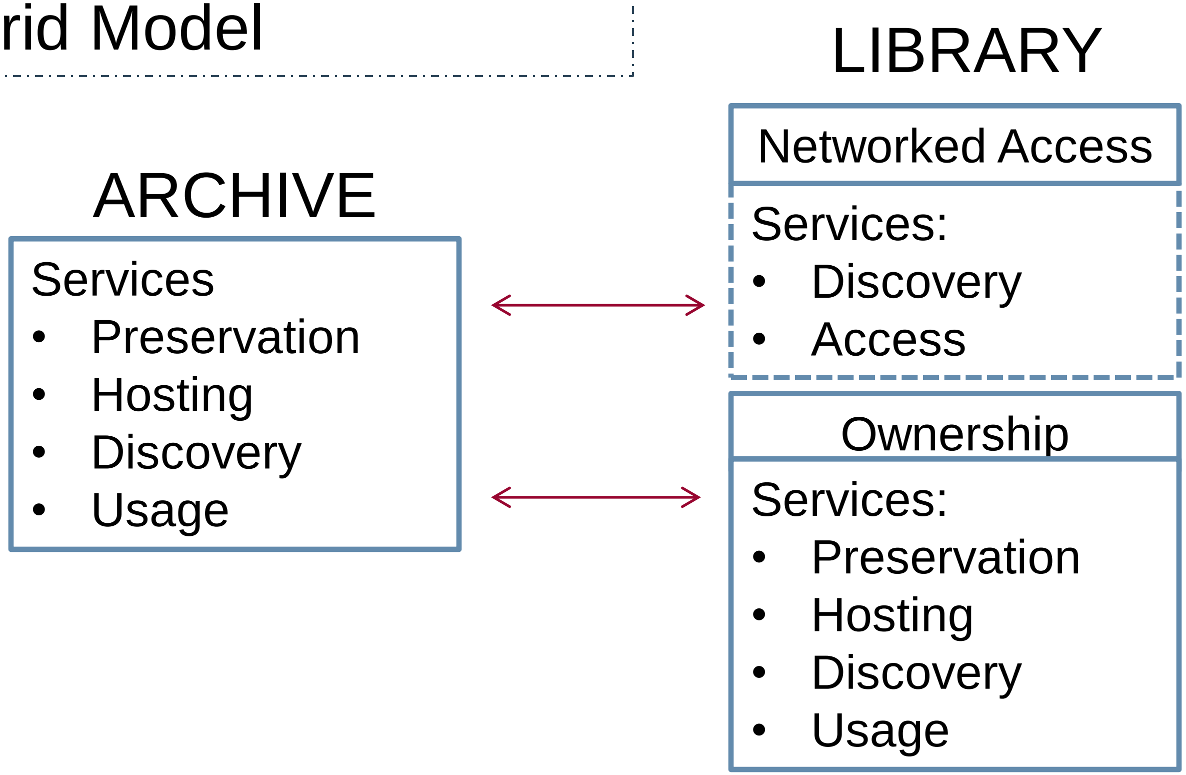
# Metrics: Focus Group Findings

- Categories
  - Scope (How much; how many)
  - Expenditures (Cost)
  - Usage (Counts)
  - Quality (Outcomes; Impacts; Value)

- Metrics that drive acquisitions
  - Retention: Cost per use
  - Selection: Usage data (when available)

# Metrics: Web Archive Service Models

1. Networked Access Model
2. Ownership Model
3. Hybrid Model



ARCHIVE

Services
- Preservation
- Hosting
- Discovery
- Usage

LIBRARY

Networked Access

Services:
- Discovery
- Access

Ownership

Services:
- Preservation
- Hosting
- Discovery
- Usage

# Metrics: Proposed Statistics
## SCOPE

- For a Web archive:
  - Size (in gigabytes, terabytes, etc.)
  - Number of discrete collections
- For each collection within a Web archive:
  - Size (in gigabytes, terabytes, etc.)
  - Number of objects by type:
    - Text
    - Image
    - Document
    - Computer file
    - Dataset
    - Video
    - Audio
    - Map

# Metrics: Proposed Statistics
## USAGE

- For each collection within a Web archive:
  - Number of sessions
    - Total number
    - Number federated or automated
  - Number of searches (queries)
    - Total number of searches run
    - Number federated or automated

# Metrics: Usage Reports

- Emulate the COUNTER usage reports for databases and journals. As such they would include:
  - Sessions by Month by Collection
  - Searches by Month by Collection
  - Searches and Sessions by Year by Collection
  - Searches and Sessions by Year by Archive
- As appropriate, these reports could be done for consortia as well as individual institution.

# Closing: Next Steps

- Subject analysis of clusters
  - Three people will evaluate each cluster (*N* = 130)
    - Identify subject terms to describe content
    - Timeframe: Summer 2011
  - Feedback to refine the cluster analysis
  - Folksonomy to describe web-published content
- Web archive metrics
  - Item Selection Profiles for SME Libraries
  - Identifying sites within EOT Archive consistent w/ profiles
- Future: Web Archive Service for the EOT Archive
  - Optimized for collection development
  - Supported by standard set of metrics